A NEW DIRECTION FOR LIBRARY DATA?

Ideas for Enabling Wider and Productive Use of Our Data
Prepared for Discussion with the StLA Steering Committee by Bob Molyneux

1. Introduction

The purpose of this paper is to outline a possible direction for future developments with library data and to solicit comment. Although I believe the ideas developed here are appropriate to all library data, I use the StLA data specifically and the NCES data series generally as a focus for the discussion because they are systematic, well-organized, and well-documented. They bespeak the enormous amount of effort that has gone into their planning and collection.

The two central problems we have with our data are a lack of a critical mass of people with an interest in analysis and the fact that anyone so disposed faces a daunting task because of the organization of these data. A key factor to improving the data is to use them and a key factor in using them is organizing them for use.

I understand that others have considered the matter of the lack of a critical mass of analytical skills in the field and have come to the conclusion I have that we need to develop a method for fostering more people interested in learning about our data and to help those interested learn suitable methods to analyze these data. In reading the "Notes" of previous StLA Steering Committee meetings, I was glad to see that you have discussed some of the same ideas. Such programs would identify potential candidates and venues for them to learn more about the data and shortcuts in using it.

We can imagine a program of internships and independent studies where people with an interest would be able to work with the people or institutions in the library data world that are involved with our data in a formal arrangement. Such a program would identify potential candidates and venues for them to learn more as well as providing opportunities for students in library schools.

Here, though, I wish to focus on the data themselves, while touching on this aspect of the data problem briefly.

This paper proposes making our published data more usable for those inclined to analyze them in the library world. The plan proposes measured stages because the task is complex. I discuss the complexity of the data next in order to make the nature of the problem before us clearer. The discussion about the data's complexity will focus first on formats and then on variables. After I discuss these points with a few tactical suggestions, strategic directions will be outlined.

As I mentioned, the major datasets I deal with here are those published by NCES. There are other sets of library data available such as those from the Association of Research Libraries (ARL) which is the longest running library data series in the world, those in digital form but proprietary such as those of the Association of College and Research Libraries (ACRL), and others of historical importance but not available such as the 1977 survey of state libraries and the COSLA data compiled by Joe Shubert in the 60s and 70s and others. Any comprehensive plan must take into account all these sets of data in appropriate ways. We should envision a system which can encompass all library data and all variables reported in any data series.

2. The Data

NCES data are collected primarily for making estimates related to national information policy and other public policy questions. The different series have varied constituencies.

Joe Shubert's December 7, 2000 paper “StLA Data and Public Policy Questions...” lists five groups that the StLA survey has been designed to supply data for. Each of the various library surveys conducted by NCES probably has similar varied constituencies. These data, then, serve many masters and to serve them better, the data must be adaptable to answer a variety of questions some of which we cannot now anticipate. The data should be organized in such a fashion that they can be rearranged readily to address new questions that arise.

The data are currently not organized for use by people in our field who have modest analytical skills nor organized for use in time series (longitudinal studies). They are stored in archival formats by year and by type of library survey. These archival data exchange formats are robust and were designed at a time when computer storage was expensive. As a result, these formats are compact, have short and cryptic variable names, and take up the minimum amount of space possible. Conditions are different these days and less compact data interchange formats are being developed that are easier to use by people who are not technically trained.

These data are used secondarily for comparison as with the NCES peer tools. These peer tools reflect an historic and central purpose that library data are used for:

primarily decision support by librarians such as budget planning, and secondarily for research on libraries.

Fortunately, the admirable quality of the NCES documentation permits us to consider rearranging the data for more flexible purposes, including pooling data across library type when necessary. For instance, suppose we wanted to know how much is spent for all libraries by state or how many volumes there are in all libraries in each congressional district or how many per capita? This week someone posed an interesting question to me: what percent of library resources are in Washington, DC? Some questions are not limited to type of library.

I will use two views of the State Library Agencies' data as examples. One view examines the format of one year's data and the second looks at the variables in three years of those data. Both of these views indicate the nature of the problems that we must address.

2.1. One year's data format

The first exhibit (I will hand these exhibits out at the meeting) is a printout of the StLA data for Alaska for 2001 as extracted from the data on the NCES Website. There are 166 numbers or characters per line. The total number of columns—if these were printed out on one line—would take 3,033 columns. Note the last two lines in this print out have 241 Rs (columns 2791 to 3033). These are imputation flags. Through a code, each one of these refers to the imputation status of one specific variable. Using these flags, imputation could be removed a variable at a time when required. Section 3.1 takes up this matter again.

Documentation on pages 9-23 of Data File: State Library Agencies Survey, Fiscal Year 2001 (at http://nces.ed.gov/pubs2003/2003342.pdf) tells how to decode the information that you see here. The State Library's name (variable name: STLANAME) begins at column 1 (the "Start @" in the second exhibit) and goes through column 63 (the end point of this variable which the documentation identifies by the variable's "length"). Columns 64-123 have PHYSADDR—the street address. And so on. For example, the number of bookmobiles is reported by BKMOBILE, columns 767-771. Suppose you want to know the average number of bookmobiles by state? You would have to find a way to get the datasets from NCES, and then parse that item out, then do the calculation.

There are two basic ways you might do that today if you downloaded the data from the NCES site and had data organized as in the first exhibit. One is to buy SAS or SPSS, learn it, and write a program which will go out to those columns and read the data in. Learning these programs takes time. Second, you might try to do it in Excel which, as you would soon discover is a bit easier said than done—particularly if you want to put multiple year's data in the spreadsheet. If you did that, then you will still have your data in Excel which is a chancy program to be manipulating data with. The point is that this first step on your quest to find the average number of bookmobiles or any other set of data is a big one.

Please understand that the NCES data have been archived in the most reasonable way possible. It is a formatting scheme that is widely used and I have stored data in similarly structure formats. It is how professionals archive data.

2.2. Three years of variables

For a second view of these data, I took data from the State Library Agencies (StLA) surveys for the years 1994, 1998, and 2001 and summarized the results of my analysis in the second exhibit. This sample was taken to get the two end years and one in the middle. The conclusions drawn are for this set of survey data only although I suspect they are broadly generalizable to all the StLA data and the other NCES library data based on my survey of the documentation. Of course, this is a first approximation and if we do create new datasets, the work done here will be revisited when the review of variables is carried out.

I used the NCES documentation to create three spreadsheets listing the variables named in each year's data, the type of the data (numeric or alphanumeric), and the starting point for the named variable in each year's data ("Start @"). I then sorted the variable names alphabetically and tried to find out how many variables were used and how each year's variables compared with the others. 1994 had 464 variables reported; 1998, 520; 2001, 668 (427 without the imputation flags). There were 347 reported in all of the three years and a total of 884 variables. "Total" is defined as one variable in any one year. The spreadsheet has the detailed data. Summary data are at the end of the spreadsheet.

Note that the "Start @" point for the variables is different each year. This fact means that a different program would have to be written to extract each year's data from the documented file.

These kinds of result are expected because of the nature of data. Data reflect the world they attempt to measure and the world changes over time. LSCA gives way to LSTA, the Internet explodes, the focus of our interest changes also, and as we learn more, we refine what we attempt to measure. This process is inevitable and healthy.

At this time, I did not attempt to see if variables with different names actually measured the same things or if variables with the same names measured different things. We will see examples of both types of errors when the closer review of these variables is conducted. These types of errors are also expected as a result of the short variable names and a focus on one year's collection at a time. Ideally, each variable name should be unique and should measure one thing across all data series and a review of the variables would have to bring this state about.

3. What steps do we take to make the data more usable?

3.1. Imputed data.

NCES StLA data have some imputed data which would have to be removed for some uses. “Imputation” results when the original survey as filled out by a representative of a library but in some cases, data were missing, made no apparent sense, as for instance, when things which should total up do not, or other anomalies. In those cases, data are generated by algorithms that are guesses of what the true but unknown number would be if it had been filled in. If your task is to make national estimates for purposes of assessing aspects of national information policy, imputation is defensible but for research purposes, it is not good practice. The 2001 format has imputed data noted with the imputation “flags” discussed above in Section 2.1. Neither the 1994 nor 1998 data have such flags. The peer tools do not use imputed data so these data do exist.

As the look at the data format in exhibit one shows, removing imputed variables would be time consuming and careful work. Ideally, we would probably need both kinds of data in a comprehensive system, depending on the purpose of the analysis.

3.2. Longitudinal data

The first step in making the data easier to use would be to create longitudinal datasets of each of the surveys. That is, all the data for each type of library for each year in one file.

Longitudinal data are useful for many kinds of analysis that are difficult under current circumstances such as following trends—they allow the analysis of libraries as dynamic, rather than static, entities. In addition, longitudinal data tend to be better behaved. If anyone wanted one year's worth of data from the longitudinal data, they could be extracted readily. What would be involved at a minimum if we are to produce longitudinal data from the StLA surveys?

As mentioned, an import program would have to be written for each year's data because the variables change their position in each file as the second exhibit shows. So, for the StLA data, that means 8 (1994-2001) input programs to write and debug.

There is a possibility, though, that the master file of these data is in a format that is easier to use. We will get to this point in Section 4.

3.3. Datasets to be produced

3.3.1. Simple datasets

I have proposed simple datasets for use by students in library schools. This idea is clearly aimed at enticing students to our data and to get practice using them. I discussed this idea with several colleagues at ALISE and was surprised by the interest it attracted. Such datasets might include, for example, the most important 10 or 20 variables included in each year of the StLA data. Or imagine richer datasets with more variables for the state libraries where, for instance, library schools are. Then students could learn to use the data to compare their state's library with others. Imagine a class exercise where students take these data and do a budget justification or try to figure out where resources could or should be allocated to make their library better.

I would envision these data being produced in a variety of formats so that students could practice with them. Excel, comma-delimited ASCII, and a format such as the raw NCES data come in, would all provide students useful practice.

3.3.2. Complex datasets

Of course, all the data must be available for researchers.

4. Newer data formats

As mentioned, the standard type of format used by NCES was developed during a time when data storage was expensive but now it is no longer and there are data exchange formats that take this fact into account.

It would be ideal if the first time someone wanted to follow the course of state libraries over the years, he or she did not have to write the 8 input programs, debug them, regularize the variable names, get out the imputation, and just, well, use the data?

The best kinds of problems are those already solved and this is an almost-solved problem.

As it turns out, the master copies of the data appear to be SAS datasets. I talked to Elaine Kroe about this point and that is what I understood her to say. Lo! I now have the StLA and public library data in SAS datasets. SAS (http://www.sas.com/) is a statistical utility used by the folks at Census and NCES who work with these data. SAS datasets are those organized by the SAS programs for use in SAS and these datasets make using data substantially easier than using formats like the NCES format discussed in 2.1 above. SPSS (http://www.spss.com/) is another competing statistical utility and it, too, has a similar internal data format. Keith Lance, for instance, uses SPSS so we have people in our community that use this utility.

SAS (and, I presume, SPSS) datasets have the data along with information about the data such as the variable names, type of data, and so forth. When using a SAS dataset, you would not write a program to extract the data from the matrix of columnar data but, rather, write a program which picks the dataset and asks the program to calculate the mean number of “bkmobiles.” You can manipulate the data directly.

Manipulating directly is better than writing all those input programs on the raw data, of course, and it is about the state of the art right now. Alas, SAS is expensive and learning it is not easy because used properly, programming is involved. SPSS may be cheaper but it has to be learned, too. If you use SPSS, you would have to convert the NCES SAS dataset to an SPSS dataset. The SPSS folks say this is possible and I have converted some of the SAS datasets of data I have compiled to SPSS formats but not all. This is an empirical question and I am doing this experiment now on the data Elaine gave me. So far, it looks promising. I expect it to work and hope to have results of this experiment by the meeting.

Both utilities have interfaces which permit pointing and clicking while hiding the actual code from the user so they can be used in situations where pointing and clicking is good enough. But this is a practice that should be avoided for all but the simplest operations.

Excel also can be used as an exchange format and for manipulation of the data although it is limited in scope, it can be used to analyze these data. Producing subsets of master files in Excel spreadsheets is a requirement of any solution to the library data problem.

But, if the master copy of these data were in one file, then it becomes relatively simple to create subsets of the data and to answer questions relatively quickly. Producing a longitudinal dataset becomes a trivial programming exercise, although producing simple datasets still would involve a detailed review of variables as begun in the second exhibit. Ironically, producing the simple datasets is a more complex exercise than producing a complex dataset.

However, what we are working towards, is a more generalizable form for storing data and one that does not involve expensive and hard to use programs like SAS or SPSS. It must work for ordinary people who have a question that data might help answer.

Before going on to the future, consider the ARL Statistics Interactive Edition at: http://fisher.lib.virginia.edu/arl/index.html.

Here the data have been arranged in such a fashion that one can query them interactively and produce graphs and analysis. This technology exists now and last year, in another context, I asked if I could have the programs that run this site to adapt to another set of library data and the answer was that I could. Imagine all the NCES data available on a similar Website.

But....

5. The future

Where this is all headed, though, is XML. The eXtensible Markup Language would permit us to store library data in such a format that it would not be necessary to buy or learn to use SAS or SPSS and it would be possible to have programs that would permit people interested in analyzing some problem to extract whatever authoritative data they want and put move them into any format they chose without having to worry about parsing or extraction programs. I think the production of open source (that is, free) utilities to do just these things would be an essential part of any such plan.

The great advantage of the ARL Interactive Edition is that it permits its users to analyze key questions from that universe's data readily and we could do well to emulate it. Its disadvantage is that it does not permit the analysis of questions that fall outside the design parameters of this Interactive Edition.

5.1. A bit about XML

There are many books written about XML so this is just a simplified description for now. It is an order of magnitude more complex than what has been discussed so far and this is not intended on being more than an introduction.

XML is an outgrowth of the development of markup languages such as SGML (the Standard Generalized Markup Language) and more recently the HyperText Markup Language (HTML), the code that Internet Web pages are written in.

In both of these languages, the tags used for markup are formally specified so, for instance, Headers can be used to markup text. In order to fool *Word*, I am going to slightly alter the structure of the tags here but you would encode a major heading roughly like this:
< H1 >Major Headline< /H1 > and a headline under it like this:
< H2 >Minor Headline< /H2>. H1 is more important than H2.

You would not see this code in the Web page unless you reveal the codes as you can do in your browser.

What XML allows is for the user to specify a formal language with its own tags. For instance, suppose we had an XML library data exchange format. We might have these tags:
< Expenditures_for_library_materials >123,456<
/Expenditures_for_library_materials > or < Bookmobiles >5< /Bookmobiles >.

Now, you would not see the tags if you were actually manipulating the data. These tags would be used to describe the data and much like in Excel or SAS datasets, you would not normally look into the storage format. Programs would do that, extract the data you need (parse out all the numbers surrounded by the “Bookmobiles” tag, for instance), and put it in a format you wanted.

You would ask the program to fetch this or that set of data and give them you to in a PowerPoint slide or a spreadsheet. Setting this all up is quite complex but the result would be the kind of thing you see with the ARL interactive Web site only more flexible and one that could address questions its developers had not thought of by people without a degree in statistical computing.

What do I mean by “flexible?” I discussed the changeable nature of data and there is also the changeable nature of queries people make on data. New questions arise and insights from analysis cause new questions. Change, then, is one constant. As a conceptual question: should one design a system by trying to anticipate all questions to be put to it or by trying to make it flexible so that it can be used to answer questions we can’t anticipate today? Properly done, XML will permit a flexible data system that does not preclude the production of the kind of Web site that ARL has provided.

XML is the ultimate direction and work is going on now with creating an XML infrastructure for many types of data, including library data. There is much work that goes before us that we can adapt. In the library world, there are examples:

* Z39.83, the NISO standard for exchanging information from circulation systems (http://www.niso.org/standards/standard_detail.cfm?std_id=728) has a:

Protocol Implementation Profile available at:
http://www.niso.org/standards/standard_detail.cfm?std_id=805 that includes the XML infrastructure.

* Z39.7-2002 is the draft NISO standard for Library Statistics and it does not yet have an XML infrastructure but ultimately will be the anchor around which such a comprehensive XML library data standard will be build.

* ONIX, at http://www.editeur.org/ is doing work in XML, including an XML standard for exchanging serials information.

* Project COUNTER is an organization comprised of people from publishers, vendors, and libraries. It has recently released its Code of Practice which is the start of a standard method for the exchange of data on the use of online materials. It will

be developing an XML infrastructure which it plans on having done this year. (http://www.projectcounter.org/)

* Roy Tennant's *XML in Libraries* (Neal-Shuman: 2002; ISBN: 1555704433) was published in 2002 and reviews other library applications of XML.

5.2. What will it take to devise such a comprehensive XML standard?

First of all, nothing suggested here contradicts the ultimate development of such a standard. The data can be read in with the NCES formats or SAS or whatever. Variables will still have to be reviewed and defined and most programs such as SAS, SPSS, Cold Fusion, Excel are being developed to handle XML.

The various library XML standards will be studied so that nothing done with these data conflict with any of those. For years one could criticize Z39.7 as a not very useful exercise but the new guidelines are impressive and provide a framework to focus the development of an XML infrastructure to accommodate our data. I said elsewhere it will be the anchor around which the work discussed in this memo will be done.

Using Z39.7, variable names can be translated from the short, cryptic names developed at a time when space was expensive and at a premium to descriptive names that are used uniquely for each variable appearing in any set of library data. While this work is going on, prototyping of extraction programs will be carried on. The initial planning process will detail the steps and the most sensible order to take them in.

5.3. Summary

What is proposed here then is the development a comprehensive and systematic plan to encompass all variables, from all open source surveys, for all years, from all types of libraries and other information agencies, in one system based, ultimately, on XML or whatever is developed from it.

This work will not be easy. It will be detailed and painstaking and it will some require skills that are unusual in our world but we already have many of the pieces in place—for example the exemplary NCES documentation. The result of this work will, however, be to transform our data from something that we collect but rarely use to something that can be used by anyone with a question about how libraries work or how better to fund them.

It is the correct direction for our data and a worthy challenge.